

Developing a Standard for Meta-Descriptions of Multimedia Language Resources

Daan Broeder⁺, Pirkko Suihkonen⁺⁺, Peter Wittenburg⁺
⁺MPI for Psycholinguistics, ⁺⁺MPI for Evolutionary Anthropology

Max-Panck Institute for Psycholinguistics
Wundtlaan 1
6525 XD Nijmegen
The Netherlands

Email: d.g.broeder@mpi.nl

Abstract:

This paper wants to discuss a standard for meta-descriptions that can be used to easy find and locate suitable multimedia/multimodal language resources (MMLR) in the Internet. We try to reach a consensus within a representative part of the linguistic community about a standard for such meta descriptions. A machine readable implementation of this standard will then allow us to build up a searchable and browsable space. Our presentation is based on the work executed within the framework of the international EAGLES/ISLE project, on practical work with meta descriptions at the MPI for Psycholinguistics and on suggestions with respect to meta data within the DOBES project.

Developing a Standard for Meta-Descriptions of Multimedia Language Resources

Daan Broeder⁺, Pirkko Suihkonen⁺⁺, Peter Wittenburg⁺
⁺MPI for Psycholinguistics, ⁺⁺MPI for Evolutionary Anthropology

Introduction

Two years ago at the Max-Planck Institute for Psycholinguistics (MPI) we started looking for a way to organise and describe the language resources produced by our researchers. This work resulted in the Browsable Corpus (BC) metadata set and a browser and metadata editor tool that work with this set [1], [2]. It also led us to participate in the ISLE/EAGLES project [3] where the MPI is responsible for the metadata part. The IMDI proposal [4] that resulted from our work in this project is at the basis of the paper presented here. The BC and the IMDI proposals are also at the basis of the software and infrastructure being developed for the DOBES project [5], [6].

In this paper we try to place the effort of creating a standard for metadata for language resources (LR) in a wider perspective, identify special requirements that the linguistic domain demands and describe the IMDI proposal.

Ideally when developing a metadata set one first discusses the vocabulary and logical structure needed to describe the domain and when agreement is reached turns to metadata syntax and implementation issues. In this paper we also refer to some implementation issues because we think that some of these issues influence the design specifications.

We refer also to the implementation done for some projects because we feel that the results reached there can stimulate further development and acceptance of the concept of metadata for language resources.

There is a glossary of terms at the end of the paper.

1 Metadata for Linguistic Resources

Until recently the use of the name “metadata” was reserved to the librarian community where metadata records are used for resource discovery of and information exchange about library resources. With the advent of the Internet and cheaper and efficient means

to digitise information more and more communities can have their particular information resources on-line on the Internet. They must consider ways for making this information available for search and retrieval in an organised and structured way, i.e. they need to design a metadata vocabulary and structure that enables others to locate specific resources. So it is also for the linguistic community, it is encouraging that after a somewhat lukewarm reception of the idea of a standard for metadata for language resources there seems to be almost universal agreement on the need of a standard.

In fact linguists have been using metadata for a long time. They just called it by other names such as “header data” after the place it was stored i.e. the header of a transcription or annotation file or the metadata was merged with the content as with the TEI [7]. In these cases the metadata and content data were contained in the same file. It is advantageous though, as will be explained later to put content and metadata in separate containers (see 5 Accessibility).

One of the fundamental issues with the design of a metadata set is the choice between a minimal metadata vocabulary and a specialised exhaustive one. Choosing a minimal one has as advantage that less work has to be done when providing metadata for a resource and it will be easier for other communities and their tools to deal with a limited and more general set of elements. A disadvantage of a minimal set is of course that only general questions can be answered. Using an exhaustive set requires providing much data and because the metadata elements get to be very specialised they become incomprehensible for other communities. In the IMDI proposal we are more inclined towards the exhaustive approach because we feel that the disadvantages associated with this approach may be overcome.

First of all the necessity for the data provider to provide many different metadata elements can be remedied for a large part by developing special metadata editors. It should be taken into account that in most projects the metadata for different resources vary only in a few fields. A metadata editor allows the data provider to use existing metadata descriptions to generate new ones. This will considerably reduce the amount of typing involved. The second problem is the lack of interoperability between a specialised exhaustive metadata set and other communities and their tools. Here a solution could be a scheme where resources are tagged by parallel metadata sets. This is suggested by the Open Archives Initiative (OAI) [8] where a specialised metadata set is accompanied by a general non specialised one. It requires though a partial mapping between the two. At the moment the possibility of a partial mapping between the IMDI set and the general Dublin-Core [10] set (as favored by OAI) is under investigation.

2 Requirements for Search and Browsing

Traditionally metadata has been used for resource discovery only. In the linguistic domain we would like to use metadata also for browsing corpora structures as it is used for instance in ICE [9] and BC. In fact we would like to use it for building and browsing

a super corpus structure that we in the BC project have named the “Browsable Corpus Universe”.

In this Browsable Corpus Universe all corpora are seen as built up from sub corpora that share certain characteristics or metadata elements. The corpora themselves form the sub corpora of the BC universe

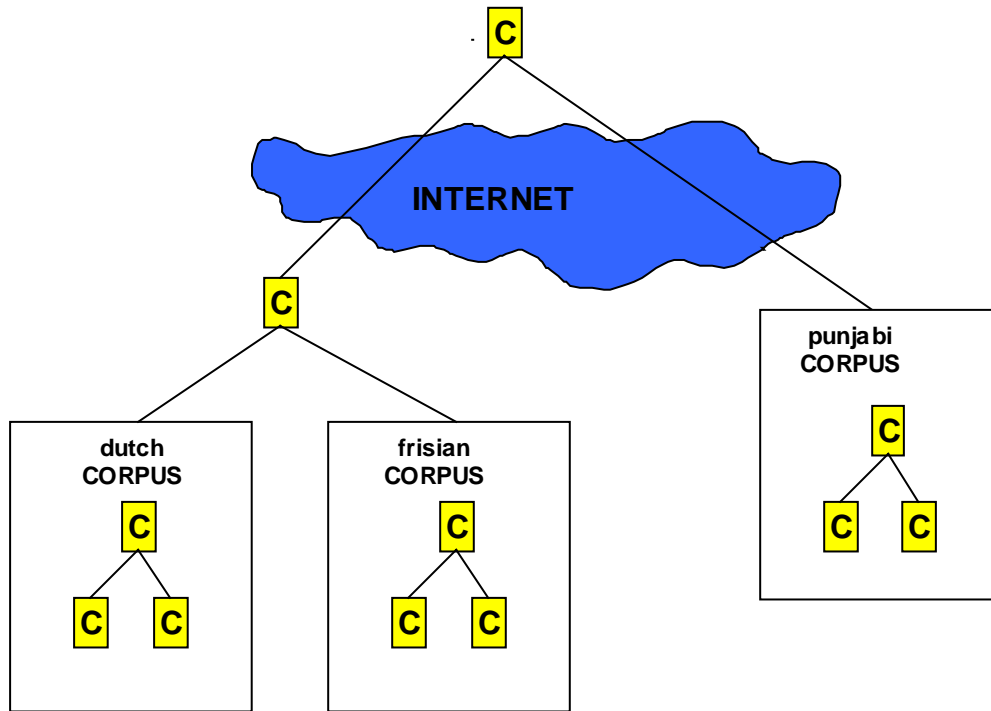


Figure 1 The Browsable Corpus Universe

In Figure 1 there is a schematic example of part of such a superstructure. There are three corpora, the Dutch and Frisian corpora are present at one site and are bundled together to form a “super corpus”. This super corpus is bundled together with a Punjabi corpus at another (physical) site to form a new corpus. This example only serves to emphasize the distributed nature of the Browsable Corpus Universe. The structure presented in this form may be uninteresting. It is possible though to have several parallel browsing structures available.

This wish to browse both at macro and micro level poses the requirement that a metadata set for the linguistic domain contains elements that can describe corpora as a whole at the catalogue level as well as elements that allow finding individual resources within corpora. We would like to be able to answer not only the question “find me all corpora with yaminjung speakers” but also “find me all recordings with female yaminjung speakers younger than 25”. To be able to answer specialised questions like this we need a large specialised metadata vocabulary. We cannot use general metadata schema like for

instance the Dublin-Core (DC) schema [10] which was designed and used for library resources. This even if we use the DC extension scheme that uses qualifiers to specialise the semantics of their metadata elements [11].

The ability to browse depends on the availability of “human readable” descriptions of (sub-) corpora and resources. Therefore we have provided the possibility to specify or link in (URL) references to such descriptions at many levels in the IMDI metadata set.

Although the Inter- and Intranet provides us with a powerful way of getting to our resources we should take care not to exclude local access i.e. a user that has stored resources on his home PC will want to be able to have the same opportunities with respect to search and retrieval for those local resources. Therefore our metadata set should support local infrastructure as well as an Internet based one. It should then also be possible to copy part of a set of resources from Internet storage to local storage. This will improve distributed control over the resources and allows reusability of parts of the corpus tree without great effort. In the BC and DOBES projects we work with a system where metadata records are stored in so called meta-description files (MDF). All tools that work with these files are able to access them either locally via the file system or over the Internet via an HTTP server.

3 Language Resource Clustering

The standard way of associating metadata with resources is resource centric. There is one resource complete in itself and usually specified by an URL that is accompanied by a bundle of metadata.

In our linguistic domain the situation is slightly different in the sense that although the linguistic resources we deal with can be considered independently, usually there exist clusters of related resources. For instance an anthropologist makes a video recording of an informant who describes a picture sequence. He also takes a few pictures of the informant, her family and her house. The picture description is not long after transcribed and analysed. From such an (linguistic) event that we will call a **session** there results a number of related linguistic resources:

1. Video tape
2. Photographs
3. Digitised video file
4. Digitised photographs
5. Digitisations of the images used as stimuli
6. One electronic transcription file
7. One or more electronic analysis files
8. Field notes and experiment descriptions (in electronic form)

We can think of metadata elements that are associated with just one of these resources such as for example the format of the transcription file but more often the metadata is associated with the whole session. The metadata associated with the whole session and

the metadata associated with individual resources together with references to the resources themselves are put into a meta-description file that is used as building block of our corpus infrastructure (see Figure 2.)

meta-description file

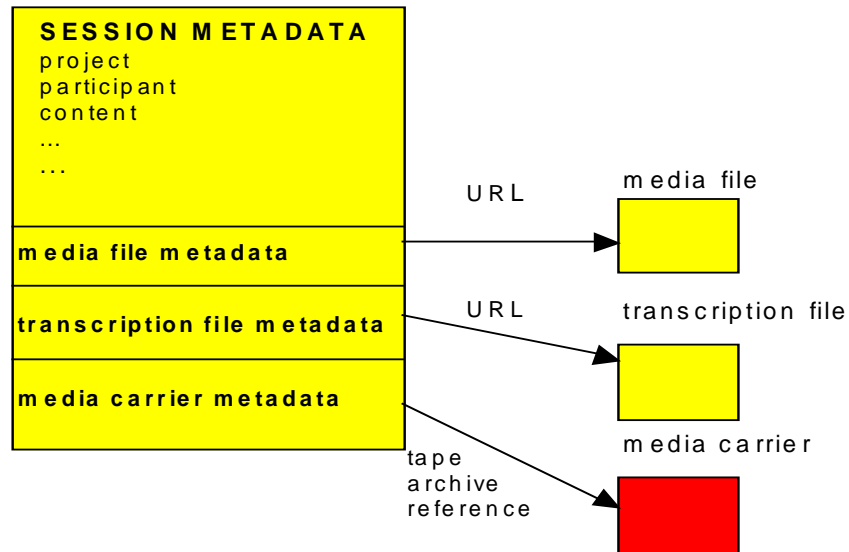


Figure 2 Meta-Description file

Our guide to designing a metadata set for linguistic resources has been finding an adequate description for this session concept. Further details may be found in “7 Metadata Structure for a Session”.

4 Tools

We distinguish between tools specially developed to work with the metadata set such as browsers and search engines and a group of presentation and analysis tools that work directly with resources. In general a user will first use such a browser or search tool to locate sessions of which the metadata meets his requirements. After that he wants to process the resources associated with the sessions with tools such as special viewers.

Some of the tools that users are likely to use work on single resources such as an image viewer that is used to visualise a digitised photograph, others work on a subset of the resources of a session such as a viewer that shows a transcription together with the audio signal in a synchronous fashion. The point we want to make here is that often a resource in itself is not sufficient to enable a tool to work. It needs extra information such as the knowledge that a specific audio file accompanies a transcription file. This information can be provided by session metadata.

For the projects the MPI is involved in (see Introduction) we have devised as a central tool a browser that is able to navigate a universe of linked meta-description files with metadata. This tool enables the execution of various tools directly after selecting

resources in the browser by using the metadata just as described above. Often users don't know and understand the limitations of the exploitation and analysis tools, which has to do with the specialization of the resources looked at. Session metadata can guide a browser in offering the user a palette of tools to work with on a specific session or resource. A good metadata set can provide information to tools that the tool cannot derive from the content data itself.

A necessary ingredient for making this tool/metadata scheme work is an appropriate coding scheme for types of resources. For the MPI's BC browser we use a private extension of the mime-type scheme. For instance a transcription file in CHAT format has as code "text/x-chat".

5 Accessibility

The question of accessibility to language resources is an important one and many working in the corpora field have noted this. For instance in the domains working with anthropological or patient data it is paramount to have access to resources barred from unauthorized persons. For practical reasons we make a distinction between access to metadata and access to the resources themselves. We would like to have the access to metadata completely free though we accept that making biographical data anonymous may be necessary. Resources themselves should be accessible via a user validation system if possible. We would like to emphasize the possibility of offering resources like transcription files in two versions; one as an anonymised version that has no access restrictions and one version with the original names with restricted access. Also possible is a scheme, where a file with information that enables to resolve codes into real names accompanies anonymised transcripts. This file is then considered as a separate resource with applicable access restrictions. It is clear that the use of anonymisation with video and audio recordings is not solved yet.

In the IMDI metadata set every pointer to a resource is accompanied by an ACCESS structure that informs on possible access restrictions and gives information on how to gain access later if this is not immediately possible. In the structure describing an informant it is possible to indicate that the name used in "full name" field is a code used to render the resource anonymous. This can be accompanied by a reference to a "protected" resource that allows resolving the codes into the real names.

Note such a simple protection scheme is only possible through the separation of metadata and content data.

6 Flexibility

Flexibility is a sore point when designing metadata sets. Too much of it and interoperability is compromised and your tools can no longer work with specialised versions of the metadata set. Too little and the use of your metadata scheme is hampered

because user cannot describe the things they want. Our guiding principle is to allow sub-communities and special projects the flexibility they need to interoperate with each other by two means:

1. Keyword/value pairs defined freely at several levels in the metadata structure
2. Metadata standard version numbering, in fact defining a new metadata set and advertising tools by means of a version number.

Users needing just a small deviation from the standard elements will use the first possibility. For instance there is no standard element for the “profession” of the informant. If such an element is needed by a project it can be added at will. The meta-description editor tool we have created at the MPI to let users generate BC format meta-descriptions makes this very simple. An interesting option would be to work with a “recommended” keyword list that could be defined per sub-community or project and is supported by the tools.

The second possibility is more rigorous. It assumes that the tools working with metadata for linguistic resources can work with more than one metadata set definition. It is likely that a metadata schema for lexica would be very different from the IMDI metadata set draft proposal as it stands. Instead of trying to force the IMDI proposal in such a way that it can accommodate lexica it is a better idea to have two different metadata sets with tools that can handle both.

7 Metadata Structure for a Session

The metadata information for a session is partitioned in a number of substructures being: “CONTENT”, “PARTICIPANTS” and “RESOURCES”. At the top level the structure contains general information such as “Name”, “Date”, “Project”, “Creator” etc. The “CONTENT” structure provides information like the language spoken, the nature of the linguistic interaction or “Genre” e.g. dialogue/monologue, spontaneous or prepared text, singing. The “PARTICIPANTS” structure gives biographical information on the participants, their linguistic background and describes their mutual relations. The information on the coding and location of the language resources themselves is contained in the “RESOURCES” structure. The macro structure is shown in Table 1 though not all elements are shown in the table. A full description is available in the IMDI draft proposal.

It may come as no surprise that XML is used as format for the metadata, this has become more or less the standard for this kind of work. We expect that the usefulness of XML will increase even more when we can use XML schema [12] as a syntax specification for the metadata set.

A thing to notice is that at several levels in the metadata structure it is possible to specify a list of keyword/value pairs. This is a way of providing flexibility to sub-communities

and special projects that can use these to specify their own metadata elements without need to turn to a new metadata set.

Session				
				Name
				Date
				...
	Project			
				Name
				Title
				Id
				...
	Creator			
	Content			Genre
				Modalities
				Key1 = val1 Key2 = val2 ...
			Language+	
				Name
				Id
				...
	Participants			
		Informant+		
				Name
				...
				Language
				Key1 = val1 Key2 = val2 ...
			Language+	
				Name
				Id
				...
		Interviewer+		...
		Contributory+		...
	Resources			
		Transcription file+		
		Annotation file+		
		Media file +		
		Data Carrier +		

Table 1 Macro Structure IMDI metadata set (not exhaustive)

References

- [1] The Browsible Corpus Project, web page, <http://www.mpi.nl/world/tg/lapp/browscorp.html>
- [2] Broeder, D.G., Brugman, H., Russel, A., and Wittenburg, P., (2000), A Browsible Corpus: accessing linguistic resources the easy way. LREC 2000 Workshop, Athens; article
- [3] EAGLES/ISLE Meta Data Initiative web page, <http://www.mpi.nl/ISLE>
- [4] ISLE Metadata Elements for Session Description Proposal http://www.mpi.nl/ISLE/documents/draft/ISLE_Metadata_2.0.pdf
- [5] DOBES project web page, <http://www.mpi.nl/DOBES>
- [6] Annotations, Formats and Data Types in the DOBES Project. Paper to be presented at the workshop on Web-Based Language Documentation and Description
- [7] Text Encoding Initiative (TEI), <http://www-tei.uic.edu/orgs/tei/>
- [8] Report on Open Archives Initiative Technical Committee Meeting - Ithaca NY, 7-8 September 2000
- [9] International Corpus of English (ICE-GB) <http://www.ucl.ac.uk/english-usage/ice-gb/>
- [10] Dublin Core Metadata initiative <http://purl.org/DC/>
- [11] Lagoze, C. (2000) - Accommodating Simplicity and Complexity in Metadata: Lessons from the Dublin Core Experience <http://ncstrl.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.cornell/TR2000-1801>
- [12] XML Schema web page, <http://www.w3.org/XML/Schema.html>

Glossary

Throughout this paper, we use a number of terms in relation to metadata. This list is partly from LAGOZE [11]:

1. *Language resource* – A transcription file, annotation file, media file, lexicon etc.
2. *Metadata for language resources* - We distinguish metadata from annotation data, knowing that many don't make this difference. While meta data in this context is meant to describe the whole language resource, the annotation is a time synchronous description of what is happening and is spoken during a recording
3. *Session* – A meaningful segment of a recording of a linguistic event with all its related data. The concept is used to bundle related language resources.
4. *Metadata vocabulary* – The set of elements (properties) provided by a specific metadata set.
5. *Metadata Schema* – The rules, or data model, for constructing statements in a metadata set.
6. *Metadata Set* – A “standard” for metadata that includes both a vocabulary and schema.
7. *Meta-Description file* – A bundle of metadata elements that describe a session or sub corpus